BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 18 2013, pages 2311–2319
doi:10.1093/bioinformatics/btt380

Systems biology

Advance Access publication July 2, 2013

The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations

Gabriele Lillacci and Mustafa Khammash*

ETH Zürich, Department of Biosystems Science and Engineering, Mattenstrasse 26, 4058, Basel, Switzerland

Associate Editor: Martin Bishop

ABSTRACT

Motivation: In the noisy cellular environment, stochastic fluctuations at the molecular level manifest as cell–cell variability at the population level that is quantifiable using high-throughput single-cell measurements. Such variability is rich with information about the cell's underlying gene regulatory networks, their architecture and the parameters of the biochemical reactions at their core.

Results: We report a novel method, called *Inference for Networks of Stochastic Interactions among Genes using High-Throughput data* (*INSIGHT*), for systematically combining high-throughput time-course flow cytometry measurements with computer-generated stochastic simulations of candidate gene network models to infer the network's stochastic model and all its parameters. By exploiting the mathematical relationships between experimental and simulated population histograms, *INSIGHT* achieves scalability, efficiency and accuracy while entirely avoiding approximate stochastic methods. We demonstrate our method on a synthetic gene network in bacteria and show that a detailed mechanistic model of this network can be estimated with high accuracy and high efficiency. Our method is completely general and can be used to infer models of signal-activated gene networks in any organism based solely on flow cytometry data and stochastic simulations.

Availability: A free C source code implementing the *INSIGHT* algorithm, together with test data is available from the authors.

Contact: mustafa.khammash@bsse.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 28, 2013; revised on June 2, 2013; accepted on June 27, 2013

1 INTRODUCTION

Gene regulation has been long recognized as an intrinsically stochastic process (Elowitz *et al*., 2002; Ozbudak *et al*., 2002). The possibly low copy numbers of the molecules involved, together with their random motion inside the cell, can cause genetically identical organisms to display different phenotypes based on variability alone (Weinberger *et al*., 2005). Ongoing progress in single cells measurement techniques is producing high-throughput datasets that allow for an increasingly accurate characterization of such variability.

*To whom correspondence should be addressed.

Stochastic computational models of gene regulation can be of great assistance in processing large amounts of experimental data and extracting information from it. The framework of stochastic chemical kinetics (Gillespie, 1976) has been successfully used to develop predictive models of systems in which random fluctuations are important (Arkin *et al*., 1998; Munsky *et al*., 2009; Neuert *et al*., 2013). Although models of this type cannot be solved directly, except for small-dimensional cases (Munsky and Khammash, 2006), they can be simulated exactly for however long and however many times is necessary to obtain the desired statistical information (Gillespie, 1977). However, stochastic simulations are computationally expensive; therefore, estimating distributions or computing summary statistics to a high accuracy can be prohibitive. Alternatively, one can resort to more compact approximate descriptions of the stochastic models, which are easier to simulate, but introduce additional error and are not applicable in all cases (Elf and Ehrenberg, 2003; Gillespie, 2000; Singh and Hespanha, 2011).

Computational models have proven to be a valuable tool in advancing the understanding of complex biological phenomena and in assisting the design of synthetic gene networks (Antunes and De Schutter, 2012; Elowitz and Leibler, 2000). Ideally, models should be able to recapitulate the known experimental observations on a process of interest and to generate new insights, which are in turn testable in new experiments. Even when pathway structure inference is able to identify the key players in the process under study, and their interactions (Marbach *et al*., 2012), to build a complete model, one critical challenge must be overcome: the determination of the many unknown parameters that will inevitably appear in it. These are numbers such as production and degradation rates, binding affinities and so forth, which are difficult to measure directly. Usually, the only available option is to measure other variables involved in the models, such as abundances of proteins of interest, and to use these measurements to infer the parameters indirectly using a dedicated computational procedure.

The problem of parameter inference for stochastic gene network models has attracted much interest in recent years, and a number of solutions have been proposed. In *maximum-likelihood* approaches, one tries to select the parameter values that maximize the probability that the model generates the observed experimental data. This has been achieved using approximations (Reinker *et al*., 2006), stochastic simulations (Daigle *et al*., 2012; Tian *et al*., 2007; Yuanfeng *et al*., 2010) and by direct solution of the stochastic models (Neuert *et al*., 2013). Maximum-likelihood

techniques have been demonstrated on biological examples, albeit on relatively small-dimensional models. *Bayesian methods* represent another attractive option. Unlike the maximum-likelihood ones, these algorithms can find many parameter values that are compatible with the experimental data, thereby producing not only point estimates but also confidence intervals. The issues related to computational cost, however, become even more severe in this case because Bayesian methods operate by testing a large number of *candidate* parameter values. For each candidate, many simulations are needed to determine whether the model output is consistent with the data. Consequently, the Bayesian framework has been mostly applied to approximate models (Golightly and Wilkinson, 2011; Komorowski *et al*., 2009; Zechner *et al*., 2012). Simulation-based methods have also been proposed (Toni *et al*., 2009) and applied with success (Liepe *et al*., 2012; Toni *et al*., 2012). In the present study, we extend their applicability to higher-dimensional stochastic chemical reaction networks by exploiting the properties of flow cytometry and other high-throughput datasets to significantly improve the computational efficiency of the inference process.

2 RESULTS

2.1 A metric for comparing experimental and simulated fluorescence samples

Likelihood-free Bayesian methods, collectively known under the name of *Approximate Bayesian Computation (ABC)*, can be used to infer unknown parameters in a model whenever the likelihood function is not available or too expensive to evaluate. They work by replacing the intractable likelihood computations with a *metric* that measures the distance between model simulations and experimental data. The key idea is to generate and test a large number of *candidate parameter sets*, also known as *particles*, and to keep track of the ones that yield simulations that are 'close enough' to the experimental data. The most basic ABC algorithm can be outlined as follows.

(1) Generate a particle θ^* from a *prior distribution* $p(\theta)$.
(2) Generate a simulated dataset X_{θ^*} using θ^*.
(3) Evaluate the distance $d(X_{\theta^*}, Y)$, where Y is the experimental dataset and d is a metric of choice.
(4) If $d(X_{\theta^*}, Y) \leq \epsilon$, where $\epsilon > 0$ is a tolerance, accept the particle θ^*, that is keep it as a suitable parameter set. Otherwise, discard it.
(5) Repeat until enough accepted particles are collected.

It can be shown that this procedure generates samples from an approximate *posterior distribution* of the unknown parameters, specifically the distribution $p(\theta|d(X, Y) \leq \epsilon)$ (Pritchard *et al*., 1999 and Supplementary Section S2).

Implementing an approach of this kind for stochastic models of gene regulation is challenging for two main reasons. First, to obtain a sufficient number of accepted particles to construct a good approximation of the target posterior distribution, a large number of candidates need to be generated and tested. Furthermore, many computationally expensive stochastic simulations are required to simulate the dataset to test each candidate.

In a previous study (Lillacci and Khammash, 2011), we investigated the properties of a specific metric for comparing simulated and experimental flow cytometry datasets. The latter typically contains data acquired from several tens of thousands of individual cells. The intensity of light emitted by each cell can be thought of as a sample of the fluorescence distribution at the time of acquisition. On the computational side, the stochastic gene regulation model can be used to generate simulated samples of the fluorescence levels. Given the two sets of samples, which we denote X and Y, the goal is to establish with high confidence whether the samples in X and the ones in Y came from the same distribution. In other words, if we denote F the distribution of the experimental samples and G the distribution of the simulated samples, we seek to test the null hypothesis $F = G$.

We proceed by computing the *empirical cumulative distribution functions (ECDFs)* associated with the two sets of samples, denoted $\hat{G}_X$ and $\hat{F}_Y$, respectively. We then measure their *Kolmogorov distance*, which is defined as follows:

$$d_{SM}\left(\hat{G}_X, \hat{F}_Y\right) = \left\|\hat{G}_X - \hat{F}_Y\right\|_\infty. \tag{1}$$

The subscripts S and M denote that the sets X and Y contain S and M samples, respectively. By the properties of norms, d_{SM} in (1) can be bounded as:

$$\begin{aligned} d_{SM}\left(\hat{G}_X, \hat{F}_Y\right) &\leq \left\|\hat{G}_X - F\right\|_\infty + \left\|\hat{F}_Y - F\right\|_\infty \\ &= d_S\left(\hat{G}_X, F\right) + d_M\left(\hat{F}_Y, F\right). \end{aligned} \tag{2}$$

We note that the quantities d_S and d_M in (2), which represent the Kolmogorov distances of the ECDFs $\hat{G}_X$ and $\hat{F}_Y$ from the unknown exact data CDF, are *random variables*, as their values change depending on the particular sets of samples X and Y that are observed. However, their null distribution is known. In other words, under the null hypothesis $F = G$, it is possible to calculate the probability that their value is less (or more) than a given threshold:

$$P\left\{d_S\left(\hat{G}_X, F\right) \leq \epsilon \middle| F = G\right\} = K_S(\epsilon). \tag{3}$$

K_S is called the *Kolmogorov distribution*, and even though it is not known in closed form, it can be evaluated numerically with high precision (Marsaglia *et al*., 2003). One key fact to note is that K_S does not depend of F but only on the number of samples S. Using this property, we can compute a *critical value* for d_S: under the null hypothesis, d_S will be smaller than or equal to $\epsilon^{(c)}(\alpha, S) = K_S^{-1}(1-\alpha)$ with probability $1-\alpha$. In other words, it is possible to find a threshold $\epsilon^{(c)}$ such that, if $F = G$, the Kolmogorov distance of a random ECDF $\hat{G}_X$ from F will be smaller than or equal to $\epsilon^{(c)}$ with high probability. Similarly, we can compute a critical value for the other distance: $d_M\left(\hat{F}_Y, F\right) \leq \epsilon^{(c)}(\alpha, M) = K_M^{-1}(1-\alpha)$ with probability $1-\alpha$.

In summary, under the null hypothesis, the following inequality must hold with probability at least $1 - \beta = (1-\alpha)^2$:

$$d_{SM}\left(\hat{G}_X, \hat{F}_Y\right) \leq K_S^{-1}(1-\alpha) + K_M^{-1}(1-\alpha). \tag{4}$$

The numbers of samples S and M, together with the probability β, which represents the confidence level of the test, define a *critical value* for d_{SM}. In other words, if d_{SM} is under the critical

value, then with probability at least $1-\beta$, the two sets of samples came from the same distribution. Equivalently, the distribution of the simulated fluorescence levels *matches* the distribution of the experimental fluorescence levels.

In our setting, the number of experimental samples M is fixed by the number of events in the flow cytometry dataset. The confidence level β is also fixed. Common values for β include 0.1, 0.05 and 0.01, corresponding to 90, 95 and 99% confidence. If we now fix the critical value for d_{SM} to a desired tolerance ϵ, we can solve (4) for S and find the number of stochastic simulations required to verify whether d_{SM} is below that tolerance with probability at least $1-\beta$. We thus obtain a 'critical' number of simulations $S^{(c)}$, which is the smallest integer such that:

$$K_S^{-1}(1-\alpha) \geq \epsilon - K_M^{-1}(1-\alpha). \quad (5)$$

If M is large, which is typically the case for flow cytometry and other high-throughput datasets, then $S^{(c)}$ is surprisingly small, allowing for the comparison of the two sets of samples to be carried out in a computationally efficient manner. A more extensive treatment of the properties of the Kolmogorov metric and of the derivation of $S^{(c)}$ is presented in the Supplementary Section S1.

2.2 Bayesian parameter inference of stochastic gene regulation models is feasible using flow cytometry and stochastic simulations

Equipped with a computationally efficient way of comparing experimental and simulated fluorescence samples, one can perform parameter inference within the framework of ABC. We have devised a novel method, referred to as *Inference for Networks of Stochastic Interactions among Genes using High-Throughput data* (*INSIGHT*), to select parameter values in such a way that the model-generated fluorescence distributions match the experimental ones from flow cytometry (Figure 1). Our proposed approach proceeds by iterating through multiple stages, in the following way.

(1) Initialization: set prior densities to specify the region of the parameter space to be searched; set an initial value of the distance tolerance ϵ.

(2) Find the number of simulations $S^{(c)}$ corresponding to ϵ.

(3) a. Propose a candidate parameter value.

b. Run $S^{(c)}$ stochastic simulations of the model.

c. Compute the distance d_{SM} of the model-generated fluorescence distributions from the experimental ones.

d. If $d_{SM} \leq \epsilon$, 'accept' the candidate, i.e. keep it as a plausible value, otherwise reject it.

e. Repeat until enough accepted candidates to approximate the parameter distribution are collected.

(4) Reduce ϵ and repeat from step 2 until the desired accuracy is achieved, or the model cannot fit the data any better.

Each *INSIGHT* stage is associated with a specific tolerance. In the first stage, the candidates are generated by random sampling from the prior densities. Each candidate is tested to check whether the corresponding model simulations match the data up to the tolerance of that stage. The collection of the candidates that satisfy the current matching condition, i.e. of the *accepted particles*, constitutes the output of the algorithm for the current stage. In each subsequent stage, the new candidates are generated by randomly picking values from the accepted particles of the previous stage. The candidates are now required to satisfy a stricter matching condition, corresponding to a lower tolerance. This process is repeated until the desired tolerance is achieved or the model simulations cannot match the data any better. The final result is a sequence of collections of accepted particles, associated to a sequence of decreasing tolerances. Each collection represents a set of samples from the approximate posterior density of the parameters $p(\theta|d_{SM}(X,Y) \leq \epsilon)$, where ϵ is the corresponding tolerance.

We remark that there exists a precise relationship between a value of the tolerance and the number of stochastic simulations required to determine whether the experimental and the simulated distributions match up to that tolerance. As the matching condition needs to be checked for every candidate, and owing to the high computational cost of stochastic simulations, the ability to use a small number of them to test the candidates quickly and accurately is crucial to the feasibility of the method. Further details on ABC and the development of the proposed *INSIGHT* algorithm are given in the Supplementary Section S2.

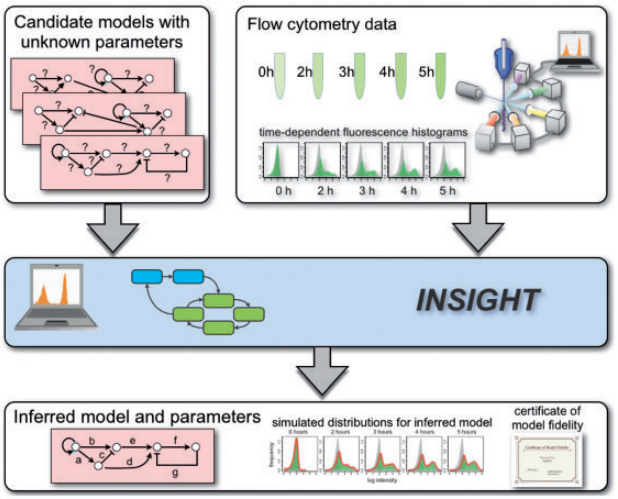

Fig. 1. The *INSIGHT* algorithm and its operation. The proposed approach uses time-dependent flow cytometry histograms to determine unknown parameters in stochastic models of gene regulation. In case of uncertain model structure, multiple candidates can be fit to the same data set and then compared in terms of their ability to fit the data. The *INSIGHT* algorithm itself comprises two nested loops. In the outer loop, a tolerance is first fixed. Owing to the properties of the Kolmogorov metric, one can then calculate how many stochastic simulations are required to compare experimental distributions and model-simulated samples up to that tolerance. In the inner loop, an approximation of the posterior density of the unknown parameters for the specified tolerance is computed. Next, the control is returned to the outer loop, where the tolerance is set to a smaller value and the process is repeated until the desired accuracy is achieved, or the model cannot fit the data any better.

2.3 A stochastic birth–death model is identified with high accuracy and high efficiency using *INSIGHT*

To demonstrate the key ideas of the method, we first applied it to a simple example, known as the *birth–death process* (Figure 2A). This stochastic chemical reaction network describes the constitutive production and degradation of a single chemical species, in this case a fluorescent protein (FP). The model keeps track of the molecular counts of the FP, which can only be changed by the random occurrences of two reactions: the creation of a new FP molecule ('birth') and the degradation of an existing FP molecule ('death'). The number of FP molecules in the system at a given time cannot be observed directly, but only indirectly through the emitted fluorescence. In analogy to Munsky *et al.* (2009), we assume that each molecule of FP emits a normally distributed random amount of fluorescence when excited by light of suitable wavelength (Supplementary Section S1.2). The mean and standard deviation of the fluorescence emitted by a single FP molecule are generally unknown. Therefore, we include them among the parameters to be estimated. This brings the total number of unknown parameters in the birth–death model to 4, which is the birth rate, the death rate and the mean and standard deviation of the unit of emitted fluorescence.

We assumed the following values for the parameters: 7.5, 0.75, 100 and 20, respectively. Using these numbers, we generated a simulated flow cytometry dataset, in which the system is observed at five different time points: 0, 0.5, 1, 2 and 10 h (Figure 2B). We then assumed that the parameters were unknown, and we applied the *INSIGHT* algorithm to identify them from the *in silico* dataset. We started from independent uniform prior densities. Figure 2C shows the joint prior density of birth rate and death rate and four joint posterior densities for four decreasing values of the tolerance. We note how, as the algorithm proceeds, the accepted particles tend to form tighter and tighter clusters around the 'true' values of the rates that were used to generate the data. This indicates that the parameters are estimated with high confidence, and that the estimates are in agreement with the true values. As a result, the fluorescence intensities generated using the medians of the posterior densities corresponding to the smallest tolerance are indistinguishable from the dataset used for the identification (Figure 2B).

2.4 Time-dependent flow cytometry measurements of the Lac-GFP system provide a rich dataset for efficient inference of a realistic stochastic model

We next applied our proposed *INSIGHT* algorithm to actual flow cytometry data. We collected fluorescence distributions from a synthetic gene regulatory network in *Escherichia coli*,

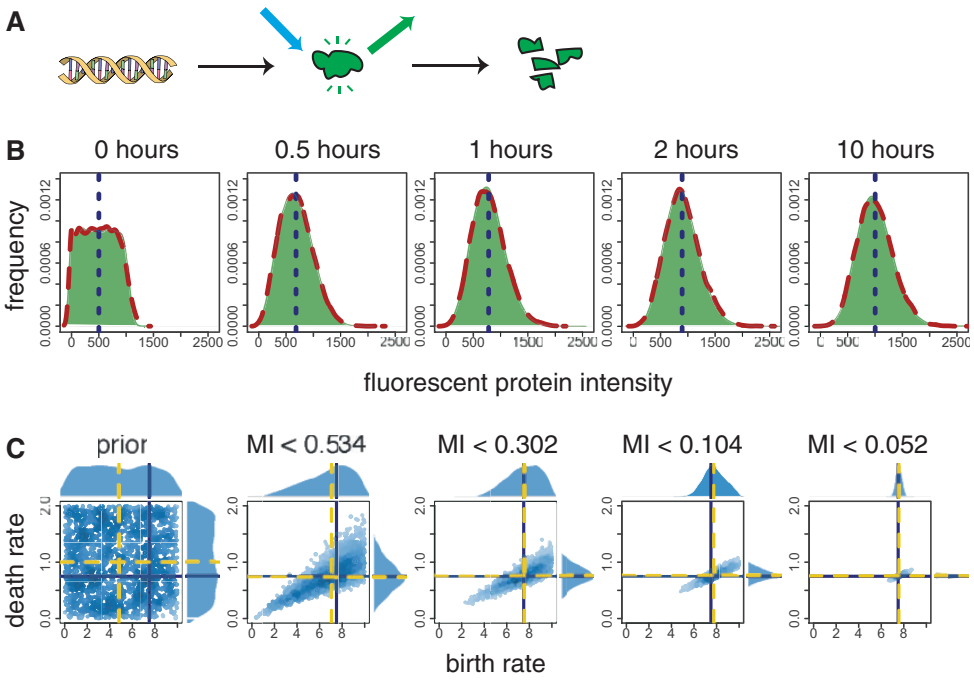

Fig. 2. Identification of a birth-death process. (**A**) Stochastic birth–death model. The birth–death process is an idealized model for the constitutive production and degradation of a single chemical species. Its molecular counts can only be changed by the random occurrence of two events: the production reaction ('birth') and the degradation reaction ('death'). (**B**) Simulated birth–death flow cytometry dataset. The time-dependent fluorescence distributions (filled) were generated using the values of 7.5 for the birth reaction rate, 0.75 for the death reaction rate and 100 and 20 for the mean and variance of the fluorescence emitted by a single protein molecule respectively. In the observed time frame, the mean fluorescence level (vertical dotted lines) shifts from 500 to 1000. The 'true' values were, then, assumed unknown, and the four parameters were estimated using the *INSIGHT* algorithm. The distributions computed using the estimated parameters (dashed lines) are almost indistinguishable from the dataset. (**C**) Evolution of the joint posterior density of the birth rate and the death rate. As the *INSIGHT* algorithm proceeds, the simulations are required to match the data with increasing accuracy. As a consequence, the densities become narrower, indicating that the unknown parameters are identified with increasing confidence. The true values of the parameters (solid lines) are well approximated by the medians of the posterior densities (dashed lines)

which we refer to as *Lac-GFP* (Figure 3A). The Lac-GFP system contains the unstable GFP variant GFP(LVA) (Andersen *et al*., 1998) under control of the Lac promoter. GFP expression is induced by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG), which inhibits the constitutively expressed repressor protein LacI. To measure and properly account for background fluorescence (Supplementary Section S1.2), we used a negative control plasmid without the gfp gene. As the control cells contain no GFP, the fluorescence that is observed when measuring them can be ascribed to autofluorescence of the host strain and to instrumental error.

The background fluorescence distribution was measured in three different experimental conditions, namely, right before induction (0 h) and 5 h after induction of the control cells with 10 μM IPTG or 100 μM IPTG. The three background distributions are practically indistinguishable, suggesting that the background fluorescence is not affected by IPTG concentration or by time (Figure 4A). In all the subsequent experiments, the 0-h distribution of the negative control cells was always used as background. GFP fluorescence distributions from the Lac-GFP system were collected right before induction (0 h), and 2, 3, 4 and 5 h after induction of the Lac-GFP cells with 10 μM IPTG (Figure 4B).

We formulated a detailed stochastic gene regulation model of the Lac-GFP system, referred to as *Lac-GFP-wt*, in which nine chemical species interact through 18 reactions (Figure 3B and Supplementary Section S3.1). The LacI repressor protein is constitutively expressed and degraded, and it can form homodimers. The dimers recognize the operator sequences on the Lac promoter (P_{lac}) and bind to them, forming a tetrameric structure that prevents transcription initiation. To account for the leakiness of P_{lac}, we allow GFP transcription from the unoccupied, partially occupied, and fully occupied P_{lac}. However, the likelihood of successful transcription initiation is highest for an unoccupied promoter, much lower for a partially occupied promoter bound to a LacI dimer, and low for a fully occupied promoter bound to a LacI tetramer. The newly synthesized GFP protein does not fluoresce right away but only after a certain maturation time. *lacI* mRNA, LacI protein, *gfp* mRNA and GFP protein are all subject to degradation. The effect of IPTG is modeled as a concentration-dependent increase in the LacI degradation rate. This is justified by the fact that our model only keeps track of functional LacI that can bind to P_{lac}. The relationship is assumed to be linear, which is of the following form: total degradation rate = basal degradation + coefficient × IPTG concentration. The mature GFP protein emits green light on excitation with blue light. Each GFP molecule is assumed to emit a normally distributed random amount of fluorescence, with fixed mean and standard deviation (to be estimated). The total number of unknown parameters in the model is 20.

The model, together with the time-dependent GFP fluorescence distributions, the background distribution and independent uniform prior densities encoding intervals of biologically plausible parameter values (Supplementary Section S3.5) were supplied to the *INSIGHT* algorithm. The method returned the posterior densities of parameter values for which the model simulations match the experimental data, up a tolerance of 0.079 (Supplementary Section S3.4 and Supplementary Figure S11). Similarly to the birth–death example, the accepted particles tend to cluster around the parameter values that, among the ones specified in the prior density, are the most likely to be compatible with the data.

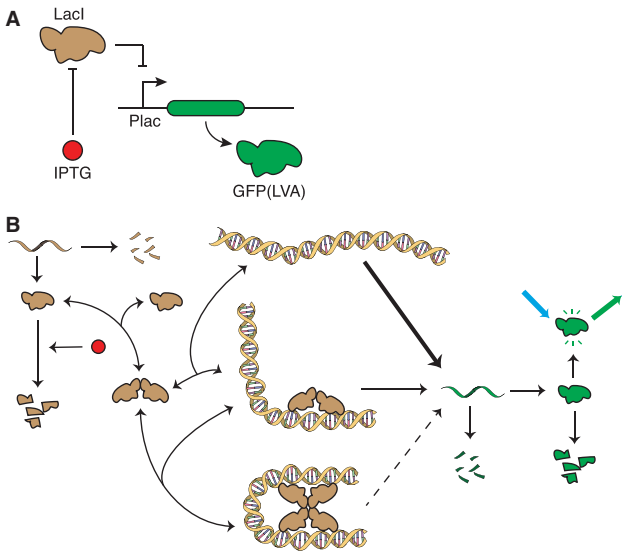

Fig. 3. The Lac-GFP system. (**A**) Gene regulation in the Lac-GFP system. In uninduced conditions, the constitutively expressed LacI repressor protein binds to the Lac promoter (P_{lac}) and inhibits transcription. The addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) inhibits LacI, thereby inducing the expression of GFP(LVA). (**B**) The *Lac-GFP-wt* stochastic model. The model is composed of 18 reactions that can randomly occur and change the molecular counts of nine interacting chemical species. The total number of unknown parameters in the model is 20

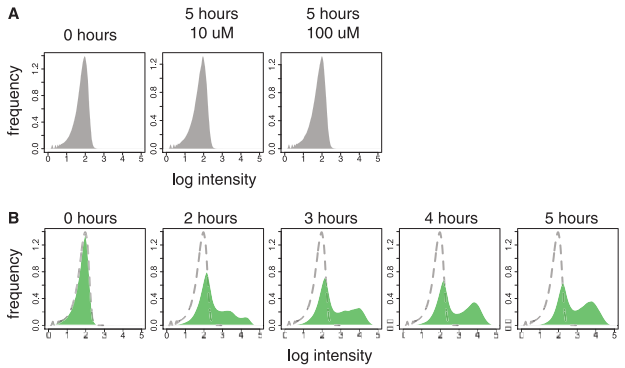

Fig. 4. Time-dependent flow cytometry measurements of the Lac-GFP system. (**A**) Background distributions measured using the negative control plasmid in different experimental conditions (filled). As no GFP is present in the control cells, the observed fluorescence can be assumed to come from the autofluorescence of the host strain, as well as from the error that is introduced by the flow cytometer. (**B**) Time-dependent GFP distributions of the Lac-GFP system. The GFP fluorescence distributions (filled) were measured from the Lac-GFP cells induced with 10 μM IPTG and collected 0, 2, 3, 4 and 5 h after induction. The distribution of the overnight culture (0 h) matches the background distribution (dashed lines) almost perfectly, indicating that no GFP was present at the beginning of the experiment

To assess the accuracy of the estimated parameters, we compared some of the 95% credible intervals of the posterior densities to values that have been previously reported in the literature. For the half-life of GFP(LVA), we obtained an estimate of (39, 67.5) min, consistent with the value of 40 min previously reported (Andersen *et al*., 1998). For the maturation time of GFP(LVA), we found a confidence interval of (3.3, 7.4) min, well in agreement with the value of 6.5 min measured before (Megerle *et al*., 2008). For the *lacI* and *gfp* mRNA half-lives, the *INSIGHT* estimates are (5.6, 10.9) min and (2.21, 5.23) min, respectively, which compare well with the median value in *E.coli* of 3.69 min that was previously found (Bernstein *et al*., 2004).

2.5 The behavior of the *INSIGHT* algorithm can reveal fundamental model inaccuracies

As aforementioned, *INSIGHT* can terminate because the model simulations cannot fit the experimental data any better. When this happens, the proposed particles are systematically rejected by the algorithm, and the inference cannot proceed. This suggests the presence of a fundamental discrepancy between experiments and model simulations. In other words, there exist no parameter values θ in the space defined by the prior densities such that the null hypothesis $F = G(\theta)$ is true.

We define *mismatch index* (MI) the quantity:

$$MI = \inf_{\theta} d(F, G(\theta)). \tag{6}$$

This number can be bounded above and below by the tolerances of the *INSIGHT* algorithm. Specifically, a tolerance for which the candidates are systematically rejected constitutes a lower bound for MI. To see this, we apply the triangle inequality twice and write:

$$\begin{aligned} d_{SM}\left(\hat{G}_X(\theta), \hat{F}_Y\right) &\leq d_M\left(F, \hat{F}_Y\right) + d_S\left(F, \hat{G}_X(\theta)\right) \\ &\leq d(F, G(\theta)) + \underbrace{d_M\left(F, \hat{F}_Y\right)}_{\phi_1} \\ &\quad + \underbrace{d_S\left(G(\theta), \hat{G}_X(\theta)\right)}_{\phi_2(\theta)}. \end{aligned}$$

which implies:

$$d(F, G(\theta)) \geq d_{SM}\left(\hat{G}_X(\theta), \hat{F}_Y\right) - \phi_1 - \phi_2(\theta).$$

Suppose now that for a given model the proposed particles are systematically rejected at tolerance ϵ_T (corresponding to the critical number of simulations $S_T^{(c)}$. This implies $d_{SM}\left(\hat{G}_X(\theta), \hat{F}_Y\right) > \epsilon_T$ for all θ, and for all $S \geq S_T^{(c)}$. Therefore,

$$d(F, G(\theta)) \geq \epsilon_T - \phi_1 - \phi_2(\theta).$$

Taking the limit as $S \to \infty$, for which $\phi_2(\theta) \to 0$ for all θ, we finally obtain:

$$MI \geq \epsilon_T - \phi_1. \tag{7}$$

In summary, in case of early termination, the tolerance ϵ_T for which the proposed particles are systematically rejected constitutes a lower bound for the MI.

By applying a similar reasoning, one can show that if ϵ_{T-1} is a tolerance corresponding to a population of accepted particles, an upper bound for MI is given by:

$$MI \leq \epsilon_{T-1} + \phi_1 + \phi_2(\theta). \tag{8}$$

Both numbers ϕ_1 and ϕ_2 can be upper-bounded (for all θ) using the properties of the Kolmogorov distance, thereby giving θ-independent bounds for MI. Further details are provided in the Supplementary Section S2.4.

We note that the MI upper bound is based on tolerances that correspond to sets of accepted particles in SMC stages. Therefore, it represents a *certificate* of model quality, in the sense that there exist parameter values for which the MI is less than or equal to the upper bound with high probability. On the other hand, the MI lower bound is based on tolerances that correspond to systematic particle rejection. As in practice, only a finite number of candidates can be checked, this may not be enough to invalidate a model, but it does show that the sampler was not able to find parameter values to certify the model. This provides evidence against it in comparison with other models that can be certified.

2.6 The MI associated with a model gives a measure of its fidelity

To demonstrate how the concept of MI constitutes a measure of model quality, we repeated the inference of the Lac-GFP system using different model variants and datasets. Specifically, we considered the following: *Lac-GFP-del*, a model in which the LacI protein is absent, and *Lac-GFP-wt*, the full model that was described earlier in the text (Figure 3B). *Lac-GFP-del* can be obtained from *Lac-GFP-wt* by removing all the species and reactions associated with LacI (Supplementary Section S3.2).

The two models were first identified using wild-type data (Figure 4B). For each model, we first obtained posterior densities for the parameter values by running the *INSIGHT* algorithm. Then, to assess the variability associated with the individual parameter values comprising the posterior densities, we randomly chose 100 accepted particles for each model, and we compared the cumulative distribution functions (CDFs) of the data with the ones of the model simulations computed with each the 100 parameter sets (Figure 5A and B, top plots). Furthermore, we compared the fluorescence intensities of the models computed using the maximum *a posteriori* probability (MAP) parameter estimates to the experimental fluorescence histograms (Figure 5A and B, bottom plots).

The wild-type model *Lac-GFP-wt* was run up to a final tolerance of 0.079, corresponding to an MI value no larger than 0.158. This indicates that its simulated distributions were able to match the experimental ones with the highest accuracy (Figure 5B). The simulated CDFs for the 100 randomly selected accepted particles are also narrowly clustered, suggesting that any of the parameter values that were picked enables to model to accurately approximate the data. Conversely, the *Lac-GFP-del* model attained a tolerance of 0.268 but was unable to achieve 0.249. The MI upper bound for this model is ~3.4-fold the one of *Lac-GFP-wt*. This indicates a poor ability of the model to match the experimental data. In this case, the simulated CDFs are narrowly clustered, but they are all far from the data CDFs,

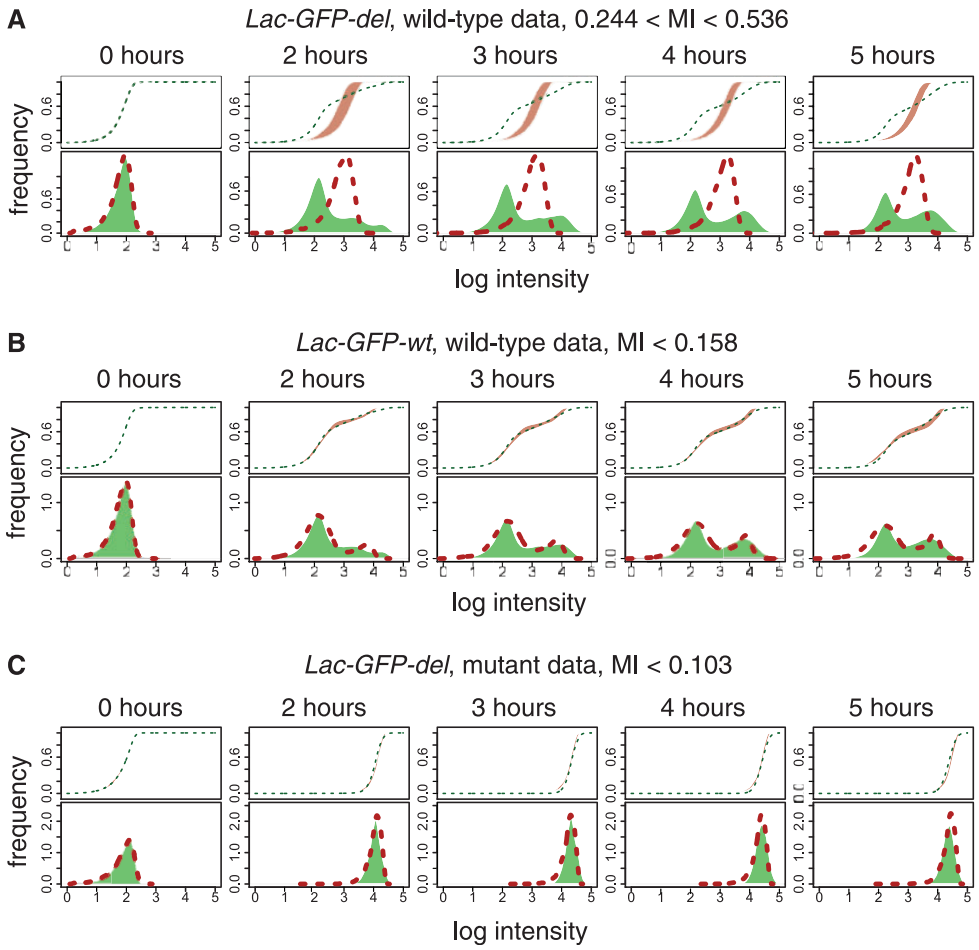

Fig. 5. Identification of different Lac-GFP models. The *INSIGHT* algorithm can give indications about fundamental discrepancies between model simulations and experimental data, measured by the *MI*. In all panels, the top plots show the time-dependent CDFs of the data (dotted lines) and 100 model-simulated CDFs for 100 different values of the estimated parameters (shaded areas). The bottom plots show the probability density functions of the data (filled) and the model-simulated probability density functions corresponding to the MAP estimates of the parameters (dashed lines). (**A**) Identification of a Lac-GFP model without the LacI protein (*Lac-GFP-del*). This model can only poorly fit the wild-type data, and all simulated distributions are far from the experimental data. (**B**) Identification of the full *Lac-GFP-wt* model. The flow cytometry distributions are fit to a high accuracy, and the MI upper bound is ~29% of the one attained by the model without LacI. (**C**) Identification of *Lac-GFP-del* using data from a *lac*-negative strain. Here, the model matches the biological process being identified. As expected, the data are fit almost perfectly, with the model simulations being indistinguishable from the data

suggesting that the performance of the *Lac-GFP-del* model is consistently poor for all accepted particles.

Furthermore, we repeated the identification of *Lac-GFP-del* using data from a *lac*-negative bacterial strain (Section 4). As the model is now matching the experimental conditions, we expected to obtain a much smaller estimate for the MI. Indeed, we ran the *INSIGHT* algorithm up to a tolerance of 0.052, associated with an MI upper bound of 0.103 (Figure 5C). The model simulations using the MAP parameter estimates are indistinguishable from the data, and the 100 simulated CDFs are narrowly clustered around the data CDFs.

As an additional control, we performed inference on a model in which the LacI repressor protein is a mutant that can not form tetramers (*Lac-GFP-mut*, Supplementary Section S3.3). Even though this model presents only a minimal structural difference from *Lac-GFP-wt*, the lack of tetramerization prevents the Lac promoter from being in a fully repressed state. This results in the model being poorly capable of fitting the wild-type data, as revealed by its high MI between 0.244 and 0.536 (Supplementary Section S3.7).

3 DISCUSSION

We described a novel Bayesian method for inference of stochastic gene regulation models using flow cytometry, referred to as *INSIGHT*. Our proposed approach uses time-dependent fluorescence distributions to find the unknown parameters in stochastic gene regulation models and does so in a way that alleviates many of the limitations of the existing techniques.

Bayesian methods represent an attractive option for estimation, as they can infer ranges of possible values for the unknown parameters, as opposed to just point estimates. However, their

use in stochastic gene regulation models has been limited due to issues of computational feasibility. By choosing the Kolmogorov metric to evaluate the distance between simulated and experimental fluorescence samples, one can calculate how many stochastic simulations are needed to determine whether such distance is above or below a certain tolerance with high probability. This fits naturally in the framework of ABC, in which metric conditions of this type are used in place of intractable likelihood functions to accept or reject candidate parameter values. The number of simulations, which we denoted $S^{(c)}$, decreases as the number of samples in the experimental dataset increases. In flow cytometry, one typically collects data from several tens (if not hundreds) of thousands of individual cells. For many practically relevant cases, this translates into a surprisingly small value of $S^{(c)}$. As an example, consider our Lac-GFP dataset, which contains 80 000 events for each time point. Comparing model-generated fluorescence samples to this dataset with a tolerance $\epsilon = 0.079$, which is enough for an accurate fit (Figure 5B), requires only 500 simulations. On the 48-core parallel machine, we used to run the inference, each parallel thread needs to run at most 11 model simulations. Therefore, each candidate parameter set can be tested in the time that is necessary to simulate the model 11 times. In other words, by using the Kolmogorov metric, it becomes possible to perform ABC inference even for a large and detailed stochastic model such as *Lac-GFP-wt*.

A distinctive feature of the *INSIGHT* method is its ability to perform inference on the exact stochastic gene regulation models. Owing to the computational cost of stochastic simulation, many approaches that have been previously reported have used different approximations to replace the stochastic models with differential equations. In some cases, these can be easier to simulate. Although techniques of this type have been used with some success (Zechner *et al*., 2012), their validity cannot be guaranteed in general, as it is difficult to assess a priori whether the approximate models reproduce the stochastic simulations faithfully. By relying only on the ability to simulate the models, this issue is circumvented entirely in the *INSIGHT* algorithm. Furthermore, as only *samples* are required to compare the simulated distributions with the experimental ones, *INSIGHT* is not subject to the model size limitations that are present in some of the approaches that work by directly solving the stochastic models (Munsky *et al*., 2009; Neuert *et al*., 2013). We also note that *INSIGHT*, like other ABC methods, can incorporate any stochastic simulation algorithm, as long as it produces independent simulations.

Another important aspect of *INSIGHT* is its ability to give indications as to which structure is the most plausible for a gene network. As the algorithm proceeds, the simulations are required to match the data to an increasing degree of accuracy defined by the decreasing values of the tolerance. We introduced the notion of MI to describe the situation in which the tolerance cannot be arbitrarily reduced due to a fundamental discrepancy between the model and the biological process. We showed how tolerance values can be used to calculate upper and lower bounds for the MI and how the MI of a model measures its performance in reproducing the experimental observations. This feature is particularly useful in biological problems in which the structure of the system under study is not fully known. In this situation, one can write several candidate models corresponding to different hypotheses and then compare them in terms of their MI. Similarly to what we found for the Lac-GFP variants, one expects that the model that can match the data with the highest fidelity represents the best description of the biological process.

Finally, we asked how significant are the computational savings of *INSIGHT* compared with a naive approach that does not take into account the properties of the Kolmogorov metric, specifically the connection between the number of events M in the experimental dataset, the value of the tolerance ϵ and the number of simulations $S^{(c)}$. We found that the theoretical number of stochastic simulations that would be required to perform inference to a similar level of accuracy is of the order of 30–50-fold the number of simulations required by *INSIGHT* (Supplementary Section S2.5). For the estimation of the *Lac-GFP-wt* model, this would be equivalent to a running time of almost 3 months versus ~2 days with *INSIGHT*.

In conclusion, we have shown how the choice of the Kolmogorov distance to compare experimental fluorescence histograms and model-generated sample paths can lead to significant computational savings in Bayesian parameter inference of stochastic gene regulation models. This effectively makes Bayesian analysis feasible in systems that were previously simply impossible to handle without resorting to approximations. The proposed *INSIGHT* algorithm is the first Bayesian method that combines the ability to handle problems of realistic size, the use of exact stochastic models, the ability to estimate them by only relying on simulations and the applicability to actual biological data. A free C source code implementing a parallel version of *INSIGHT*, together with the Lac-GFP data and stochastic models presented in this manuscript, is available from the authors.

4 MATERIALS AND METHODS

4.1 Stochastic modeling of gene expression

We model stochastic interactions among genes using the framework of stochastic chemical kinetics introduced in Gillespie, 1976. In this kind of models, one keeps track of the molecular counts of the chemical species of interest in the process under study. Under certain assumptions, the time evolution of the counts can be simulated exactly using the *Stochastic Simulation Algorithm* (Gillespie, 1977). A detailed summary of this methodology is presented in the Supplementary Section S1.

4.2 Construction of the Lac-GFP systems

The experiments for wild-type data collection were conducted using the *E.coli* strain MC4100, which is a K-12 strain containing [F' proAB $lacI_qZM15$ Tn10 (TetR)] from XL-1 Blue (Agilent Genomics). The negative control strain was obtained by transforming the plasmid pLAC33 (Warren *et al*., 2000) into the MC4100 cells. The Lac-GFP synthetic circuit was obtained by subcloning the unstable GFP variant GFP(LVA) (Andersen *et al*., 1998) between the BglII and SphI sites of pLAC33, thereby removing a part of the TetR gene. This plasmid was then transformed into the same MC4100 cells to obtain the Lac-GFP strain. The MC4100 strain, the Lac-GFP plasmid and the pLAC33 vector were gifts from Prof. David Low's laboratory (University of California at Santa Barbara, Santa Barbara, CA, USA).

The experiments for mutant data collection were performed using an *E.coli* strain that was obtained by deleting the *lac* and *ara* operons from the strain MG1655. The same Lac-GFP plasmid and pLAC33 vector were moved into the mutant strain to obtain two new strains: a new Lac-GFP strain without LacI, referred to as Lac-GFP-del, and the

matching negative control. The mutant strain was a gift from Stephanie Aoki (ETH Zürich, Basel, Switzerland).

4.3 Time-dependent flow cytometry measurements

Bacteria were grown in low-salt Luria-Bertani (LB) medium, containing 10 g/l of tryptone, 5 g/l of yeast extract and 4 g/l of sodium chloride. Antibiotics were used at the following concentrations: 100 μg/ml ampicillin, 40 μg/ml kanamycin and 12.5 μg/ml tetracycline.

Overnight cultures of the Lac-GFP, Lac-GFP-del the two negative control strains were grown in low-salt LB supplemented with 0.2% glucose (for the wild-type strains) or 1.5% glucose (for the mutant strains) for 12–16 h at 37°C with 230 rpm shaking. The purpose of the glucose was to ensure full repression of the Lac promoter, even in absence of LacI. A sample was collected from each overnight culture by diluting a small aliquot in sterile phosphate-buffered saline. These samples constituted the 0 h time point of the datasets.

The overnight Lac-GFP and Lac-GEP-del cultures were subcultured to an approximate OD_{600} of 0.05 and grown in low-salt LB with 10 μM IPTG at 37°C with 80 rpm shaking. Samples were collected 2, 3, 4 and 5 h following induction.

Similarly, the overnight negative control cultures were subcultured to an approximate OD_{600} of 0.05 with 10 μM and 100 μM IPTG. These were also grown at 37°C with 80 rpm shaking, and samples were collected after 5 h.

The collected samples were kept on ice at all times. The flow cytometry measurements were performed on a BD LSRFortessa^{TM} instrument (Becton, Dickinson and Co.). The fluidics were operated at the lowest possible flow rate so that the acquisition rate was consistently kept under 2000 events per second. The cells were excited using a 488 nm blue laser, operated at the maximum power of 100 mW. The fluorescence emission from GFP was detected using a 530/30 nm filter. For each sample, 150 000 raw events were recorded. The following voltages were used: forward scatter 644 V, side scatter 251 V, GFP 500 V.

The events were gated using the software FlowJo (Tree Star Inc.) based on the forward scatter and side scatter signals. Events that were likely to have been generated by noise in the acquisition process were identified by running samples consisting of phosphate-buffered saline only and removed.

ACKNOWLEDGEMENT

The authors thank Stephanie Aoki for providing materials and useful suggestions for the Lac-GFP experiments presented in this manuscript.

Funding: M.K. acknowledges research funding from the US National Science Foundation through Grant ECCS-0835847 and from the Human Frontier Science Program through Grant RGP0061/2011.

Conflict of Interest: none declared.

REFERENCES

Andersen,J.B. *et al.* (1998) New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl. Environ. Microbiol.*, **64**, 2240–2246.

Antunes,G. and De Schutter,E. (2012) A stochastic signaling network mediates the probabilistic induction of cerebellar long-term depression. *J. Neurosci.*, **32**, 9288–9300.

Arkin,A. *et al.* (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. *Genetics*, **149**, 1633–1648.

Bernstein,J.A. *et al.* (2004) Global analysis of *Escherichia coli* RNA degradosome function using DNA microarrays. *Proc. Natl Acad. Sci. USA*, **101**, 2758–2763.

Daigle,B.J. *et al.* (2012) Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC Bioinformatics*, **13**, 68.

Elf,J. and Ehrenberg,M. (2003) Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res.*, **13**, 2475–2484.

Elowitz,M. *et al.* (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Elowitz,M.B. and Leibler,S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338.

Gillespie,D.T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, **22**, 403–434.

Gillespie,D.T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.-US*, **81**, 2340–2361.

Gillespie,D.T. (2000) The chemical Langevin equation. *J. Chem. Phys.*, **113**, 297.

Golightly,A. and Wilkinson,D.J. (2011) Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, **1**, 807–820.

Komorowski,M. *et al.* (2009) Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, **10**, 343.

Liepe,J. *et al.* (2012) Calibrating spatio-temporal models of leukocyte dynamics against in vivo live-imaging data using approximate bayesian computation. *Integr. Biol.*, **4**, 335–345.

Lillacci,G. and Khammash,M. (2011) Model selection in stochastic chemical reaction networks using flow cytometry data. In: *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC2011)*. pp. 1680–1685. IEEE, Orlando, FL.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Marsaglia,G. *et al.* (2003) Evaluating Kolmogorov's distribution. *J. Stat. Softw.*, **8**, 1–4.

Megerle,J.A. *et al.* (2008) Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophys. J.*, **95**, 2103–2115.

Munsky,B. and Khammash,M. (2006) The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, **124**, 044104.

Munsky,B. *et al.* (2009) Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, **5**, 318.

Neuert,G. *et al.* (2013) Systematic identification of signal-activated stochastic gene regulation. *Science*, **339**, 584–587.

Ozbudak,E.M. *et al.* (2002) Regulation of noise in the expression of a single gene. *Nat Genet*, **31**, 69–73.

Pritchard,J.K. *et al.* (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.*, **16**, 1791–1798.

Reinker,S. *et al.* (2006) Parameter estimation in stochastic biochemical reactions. *IEE Proc. Syst. Biol.*, **153**, 168–178.

Singh,A. and Hespanha,J. (2011) Approximate moment dynamics for chemically reacting systems. *IEEE T. Automat. Contr.*, **56**, 414–418.

Tian,T. *et al.* (2007) Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics*, **23**, 84–91.

Toni,T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.

Toni,T. *et al.* (2012) Elucidating the in vivo phosphorylation dynamics of the erk map kinase using quantitative proteomics data and bayesian model selection. *Mol. BioSyst.*, **8**, 1921–1929.

Warren,J.W. *et al.* (2000) Construction and characterization of a highly regulable expression vector, pLAC11, and its multipurpose derivatives, pLAC22 and pLAC33. *Plasmid*, **44**, 138–151.

Weinberger,L.S. *et al.* (2005) Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell*, **122**, 169–182.

Yuanfeng,W. *et al.* (2010) Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent. *BMC Syst. Biol.*, **4**, 99.

Zechner,C. *et al.* (2012) Moment-based inference predicts bimodality in transient gene expression. *Proc. Natl Acad. Sci. USA*, **109**, 8340–8345.